The Thesis Committee for Neha Pundlik Srikanth
certifies that this is the approved version of the following thesis:

Characterizing Content Addition and Explanation Generation in Document-Level Text Simplification

Committee:

Junyi Jessy Li, Supervisor

Greg Durrett, Co-Supervisor

Characterizing Content Addition and Explanation Generation in Document-Level Text Simplification

by

Neha Pundlik Srikanth

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Computer Science

The University of Texas at Austin

August 2020

To my parents Mona and Srikanth, and my sister Maya.

Acknowledgments

I would like to give my sincerest thanks to my advisor, Jessy Li. Throughout this work, Jessy has been a constant source of patience, optimism, and encouragement. She has always made herself available whenever I needed help, and after every meeting or discussion, I was always left feeling motivated and excited about our research. Her support and advice have made this work possible, and I am forever grateful for all her guidance.

I'd also like to thank Greg Durrett for providing such valuable feedback on this work. As a student in Greg's NLP class, his excitement and commitment to his students was inspiring, and sparked my interest in NLP research.

Many thanks to Angie Beasley and Alison Norman for their support and guidance during my undergraduate career.

I am grateful for all the incredible friends I've made who have been there for me throughout my journey thus far. Their caring, support, and belief in me during my highs and lows helped me immensely.

I am incredibly blessed and fortunate to have such a loving, inspiring, and brilliant family. I'd like to thank my sister Maya for spending hours annotating data, having thoughtful conversations about this work with me, and for being one of my best friends. I'm indebted to my parents for giving me the opportunities and experiences that have made me who I am. Their determination and dedication to everything they do has been nothing short of inspirational.

Neha Pundlik Srikanth

The University of Texas at Austin

August 2020

Characterizing Content Addition and Explanation Generation in Document-Level Text Simplification

Neha Pundlik Srikanth, M.S.Comp.Sci.
The University of Texas at Austin, 2020

Supervisors: Junyi Jessy Li
Greg Durrett

Text simplification has remained an important task in computational linguistics for many years. Much of text simplification research focuses on modeling sentence simplification, addressing operations such as deletion, reordering, sentence splitting, and substitution, while research advancements in document-level simplification have been fairly limited. This work introduces a new phenomenon in document-level simplification called elaborative simplification, involving the *insertion* of content to make simplified texts easier to understand. We analyze the nature of elaborative simplification using a new corpus we collect, and illustrate its wide spectrum of contextual specificity, ranging from simple definitions to multi-step reasoning. We introduce two new modeling tasks - contextual specificity prediction and elaboration generation, and explore the capability of large scale pre-trained language models to generate a range of contextually specific elaborations.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Text simplification aims to help audiences read and understand a piece of text through a series of operations, including replacing complex words, removing difficult content, and modifying its structure, while staying faithful to its central idea and meaning. Text simplification remains an important task with beneficial social impact such as improving text accessibility for children (De Belder and Moens, 2010; Kajiwara et al., 2013), language learners (Yano et al., 1994; Petersen and Ostendorf, 2007; Pellow and Eskenazi, 2014; Paetzold, 2016), and those with language impairment (Carroll et al., 1998; Rello et al., 2013); it can also be used to improve downstream natural language processing applications, such as machine translation (Chen et al., 2012; Štajner and Popović, 2016) and summarization (Vanderwende et al., 2007; Silveira and Branco, 2012).

Much of recent work in text simplification has been confined to sentence simplification. While constraining text simplification to the sentence level has allowed for rapid advancements in techniques and capabilities of state of the art systems, the task formulation focuses on the sentence as an independent body of text, ignoring *relationships* between sentences and paragraphs, as in a document.

In this work, we begin to explore text simplification at the document-level

by introducing a novel phenomenon we call *elaborative simplification*. Elaborative simplification involves the insertion of content to make simplified text easier to comprehend. Effective elaborations must not only provide background, but they must do so in a contextual manner, adding relevant information to the surrounding text.

1.1 Contributions

This thesis presents the first data-driven study in elaborative simplification. Ultimately, we aim to answer the question: **When simplifying large bodies of text, how do people elaborate?** Through this study, we make several contributions:

1. Introduce and define a new phenomenon we call elaborative simplification at the document level by studying the Newsela corpus (Xu et al., 2015).

2. Develop a scheme to build a new corpus consisting of over 1.3K true elaborations from over 650 documents, using a combination of heuristic-based automatic extraction and human verification through crowdsourcing.

3. Conduct a multi-dimensional study of the nature of these elaborations by building a framework to analyze their contextual specificity and the type of knowledge they encode.

4. Define and establish baselines for two new modeling tasks - contextual specificity prediction of elaborations and elaboration generation.

5. Analyze the challenges that state-of-the-art large scale pre-trained language models face during elaboration generation and contextual specificity prediction.

1.2 Thesis Outline

This thesis is organized in the following manner: Chapter 2 explores work related to both text simplification and natural language understanding and generation. Chapter 3 introduces and explores the nature of elaborative simplification throughout the Newsela corpus. Chapter 4 introduces contextual specificity prediction and elaboration generation, outlining models, experimental settings, and model performance, reflecting on the performance and limitations of large-scale pre-trained language models with regards to elaborative simplification. Chapter 5 concludes our work and discusses future directions.

Chapter 2

Background

Elaborative simplification ostensibly falls only in the domain of text simplification. However, many elaborations require multi-hop reasoning, inference, commonsense reasoning, and relevant information retrieval, making it an interesting benchmark for a bevy of related tasks. In this chapter, we discuss relevant work that serves as the backdrop for this thesis. We first discuss related work in text simplification and its applications, as well as recent work in natural language understanding.

2.1 Text Simplification

Text simplification has been a long-standing task in computational linguistics. Its primary goal is to modify a source body of text to make it easier to read or understand. Text simplification has far-reaching impact across various audiences. (Mason, 1978) illustrate that splitting longer, more complex sentences into several shorter ones improved comprehension in low-literacy audiences. Simplified texts are often used for second language learning (Crossley et al., 2007). People with disabilities such as aphasia, (Carroll et al., 1999), deafness, or dyslexia may benefit from simplified text as well (Siddharthan, 2014).

In the past, sentence simplification has largely been approached through

four operations - deletion, reordering, substitution, and splitting (Alva-Manchego et al., 2020; Narayan et al., 2017; Zhao et al., 2018; Alva-Manchego et al., 2017; Nisioi et al., 2017). While effective in reducing complexity, these systems overlook an essential component of simplification – for certain audiences, the *insertion* of relevant background content can vastly improve readability. Elaborations inserted in the form of simple definitions, clarifications, or analysis specific to document context can aid comprehension and facilitate connections between content in the original text. For example, Ross et al. (1991) show that elaborative modification can help language learners' comprehension of passages.

As mentioned, most current state-of-the-art simplification systems tackle **sentence simplification** (SS), the task of simplifying a sentence while retaining its overall meaning. Zhang and Lapata (2017) approach sentence simplification via a classical encoder-decoder model trained using an RL framework. Kriz et al. (2019) address copying issues with encoder-decoder models for simplification by proposing a reranking model to promote fluency, adequacy, and simplicity, and increase diversity in candidate simplifications. Several edit-based simplification models have been proposed (Dong et al., 2019; Kumar et al., 2020), resembling how humans may approach manual simplification. While effective at generic simplification, many of these models produce simplifications for texts without adapting to different target audiences. Martin et al. (2019) develop a controllable sentence simplification system, allowing for users to tailor simplifications to a specific set of parameters, including desired output length, paraphrasing amount, and lexical complexity.

Many transformations in document-level simplification extend beyond sentences boundaries (Alva-Manchego et al., 2020; Petersen and Ostendorf, 2007; Siddharthan, 2014). Reducing the scope of text simplification to the sentence level has allowed for a plethora of new sentence simplification techniques, among which include those above. While these techniques do generate understandable,

Source Text Snippet	Sentence-Simplified Text	Target Text Snippet
But forecasters say the death toll could have been even higher, if they could not quickly track the rise of the rain-swollen Mississippi, Illinois and other rivers that spilled over their banks. A network of gauges provides that information to officials in states that are prone to flooding.	But forecasters say that the death could have been even higher , if they could not quickly track the rise of the rain-swollen Mississippi , Illinois and other rivers. A network of gauges provide that information to officials in states that are prone to flooding, and record the rise and fall of waters across the United States .	But the death toll could have been higher. Forecasters said more people could have died if there were no river gauges. **A gauge is a kind of measuring stick.** It sits in the water. The United States has 8,000 gauges to quickly track the rise of rivers.
A new standard would put more areas of the country in violation of air quality standards and place parts of the West in a tough spot between a rising baseline of ozone and stricter federal limits. Limiting pollution flowing in from Asia would require an internal treaty, said Owen Cooper, an atmospheric scientist at the Cooperative Institute for Research in Environmental Sciences in Boulder, Colorado.	A new standard would put more areas of the country in violation of the law and place parts of the West in a tough spot between the government. Limiting pollution in from Asia would need an internal treaty said that Owen Cooper.	It will not be easy to stop pollution from China, said scientist Owen Cooper. The U.S and China would have to work out a deal. **A deal between two countries is called a treaty.**
Claudia gets straight A's at one school, somewhat lower grades at her other. But as years pass and coursework gets more complex, the odds rise against her. Eventually, about 90 percent of kids living in seasonal worker housing drop out of school, according to the San Jose-based nonprofit human rights organization Human Agenda.	Claudia gets straight A's at one school , and about 90 % of kids living in seasonal worker housing drop out of school , according to the San Jose-based nonprofit human rights organization Human Agenda .	Claudia goes to two different schools each year. She gets straight A's at one. Her grades are lower at the other. **Switching between schools makes it more difficult to learn.** It's easy for kids like Claudia to fall behind. Nine out of every 10 children living in farmworker camps drop out of school, says Human Agenda.
"They don't seem to care," he said about his employers. "It's horrible how they manage us, how they talk to us, how they treat us. They don't respect us as human." The fast-food industry used to employ mostly younger people just trying to make some extra money as they went through school. Now, workers are older and depend on the work to feed families.	He said that "They don't seem to care " about his employers . It is about how they talk to us, how they don't treat us as human. They used to use younger people just trying to make some extra money as they went through school. Now , workers are older , and use the work to feed families.	"They don't seem to care," he said about his employers. "It's horrible how they manage us, how they talk to us, how they treat us. They don't respect us as human." **The fast-food workforce has changed over the years**. Restaurants used to hire mostly younger people just trying to earn pocket money. Now, workers are older. They depend on the work to feed their families.

Figure 2.1: Comparison of source text (left), sentence-by-sentence simplification (middle), and target text (right). Sentence simplifications are generated using the ACCESS (Martin et al., 2019) system. Verified elaborations in the target text are in bold and highlighted green.

effective simplifications of complex source text at the sentence level, they **cannot** be extended to document-level simplification (or other simplification systems with practical applications) by simply applying them one sentence at a time to sentences in the source text. SS systems draw context from individual sentences alone, while documents have an inherent cross-sentence structure, relating sentences and paragraphs to each other. Figure 2.1 illustrates one such state-of-the-art SS system, ACCESS (Martin et al., 2019), applied to text around verified elaborations. We see that these sentences are unable to generate text similar to explanations or elaborations, drawing content solely from the original text.

The introduction of large corpora of original-simplified document/sentence pairs has been instrumental in advancing simplification research (Zhu et al., 2010; Woodsend and Lapata, 2011; Coster and Kauchak, 2011). These corpora are built using Simple English Wikipedia (SEW), a collection of simplified English Wikipedia articles that use fewer words and simpler grammar useful for children, new language learners, or adults with language impediments (Alva-Manchego et al., 2020). SS systems trained on Simple Wikipedia utilize sentence alignments extracted from documents in English Wikipedia and Simple Wikipedia using a variety of similarity based heuristics such as cosine similarity, Wikipedia edit history, and word-level alignment (Hwang et al., 2015). While SEW serves as a useful resource for SS systems, upon manual examination, we find that elaborative simplification is less frequent in the Simple Wikipedia corpus. We hypothesize that this is largely due to the fact that articles are centered around a single specific entity, in turn causing deletion of extraneous facts and paraphrasing of essential facts to be the main operations performed. In addition, document pairs in the corpus tend to be of drastically different lengths, making elaborative simplification even more rare. Given the style and purpose of Simple Wikipedia, we find it to unfit for our study of elaborative simplification.

In 2015, Xu et al. (2015) introduced the Newsela corpus, highlighting several

shortcomings of the Simple Wikipedia corpus, including poor generalization to other genres, inadequate simplifications, and its tendency to be prone to sentence alignment errors. The Newsela corpus consists of 1,130 sets of articles. Each set of articles consists of 4 or 5 articles about the same event or topic manually simplified by professional editors and written for varying grade levels, ranging from grade 3 to grade 12. So far, this corpus has been primarily used to train SS systems; systems such as Zhang and Lapata (2017) rely on alignment heuristics to extract sentence pairs from article sets. However, this corpus provides a promising starting point to study document-level simplification. News articles in the corpus provide a more natural document structure than Wikipedia, reflecting potential practical settings where document simplification systems could be useful. In addition, the document length disparity between parallel articles is much less than in Simple Wikipedia. Our elaboration corpus we introduce in Chapter 3 is built from identified elaborations across 679 documents in the Newsela corpus.

2.2 Document Level Approaches

Prior document-level work has remained largely limited. Mandya et al. (2014) develop a sentence simplification system that optimizes global text constraints such as lexical density, the ratio of difficult words, and length of text. Similarly, Woodsend and Lapata (2011) globally optimize other aspects of content and style for their SS system. Both of these systems operate on the Simple Wikipedia corpus. Zhong et al. (2019) conduct the first data-driven study on sentence deletion in the Newsela corpus, analyzing discourse-level factors for sentence deletion across documents written for elementary and middle schoolers.

2.3 Natural Language Understanding and Generation

Elaborative simplification is primarily a generation task, requiring some degree of understanding and reasoning over relevant content in the original document using implicit background knowledge. Some elaborations entirely consist of making this implicit background knowledge explicit. This overlaps with areas of natural language understanding (NLU) such as commonsense reasoning (Sakaguchi et al., 2019) and reading comprehension (Rajpurkar et al., 2016).

Many successful state-of-the-art models for NLU tasks rely heavily on the background knowledge and commonsense inherently baked into large scale pre-trained language models (Davison et al., 2019). For example, Liu et al. (2019) achieved state of the art on the Winogrande dataset (Sakaguchi et al., 2019), a dataset involving expert-crafted pronoun resolution problems in the style of the Winograd Schema Challenge (Levesque et al., 2012). Petroni et al. (2019) show that without finetuning, BERT (Devlin et al., 2018) contains knowledge competitive with traditional NLP methods that have access to an oracle knowledge base. However, whether or not these large scale pre-trained language models can perform multi-step reasoning using this implicit background knowledge is still unclear. Shwartz et al. (2020) present methods to further elicit this background knowledge by intermediately asking information-seeking questions. We investigate the ability of large-scale pre-trained language models to generate elaborations involving knowledge retrieval, commonsense reasoning, and other reasoning skills by establishing baselines using GPT-2 (Radford et al., 2019), a large scale transformer-based language model trained on millions of documents scraped from the web.

Generating elaborations boils down to generating coherent sentences containing new and informative content relevant to document context, using the same language as in simplified document corpora. Part of this could involve retrieving relevant content implicitly (through techniques relying on the knowledge

embedded in large scale pre-trained LMs) or explicitly (through web scraping-like techniques). Kang et al. (2019) introduce the related task of generating *contextually relevant* entity post-modifiers. Their task is a largely constrained version of elaboration generation, framed as a data-to-text generation problem using input data from Wikidata. However, it captures the spirit of elaboration generation – ideally, we *insert* information during or after simplification of documents that is relevant to document context.

2.4 Summary

In this chapter, we surveyed relevant work in both text simplification and natural language understanding/generation, laying the groundwork for the introduction of elaborative simplification. We discussed existing document-level simplification analyses, as well as related tasks in contextual text generation.

Chapter 3

Elaborative Simplification

3.1 Chapter Overview

In this chapter, we introduce elaborative simplification, a phenomenon we observe to occur during document simplification. We discuss the construction of our new annotated dataset of 1.3K elaborations through a combination of heuristic-based automatic extraction, and expert and crowdsource annotation. We propose a contextual specificity framework to analyze the manner in which content is elaborated, and present a qualitative analysis of elaborations themselves, including the range of knowledge they capture.

3.2 Phenomenon Overview

Elaborative simplification involves the insertion of content in the form of definitions, details, clarifications, or analyses, to provide readers with necessary additional context and improve readability of simplified text. We consider a sentence an elaboration if it contains new content (such as entities, actions, descriptions, or concepts) present in the simplified document, but semantically missing from the original document. Note that elaborations may contain multiple sentences, but we

Original Text
Results, she said, "could help the team better understand ancient Egyptian health" and, correspondingly, modern-day health. For instance, some mummies still have arteries in their mummified remains, Miller-Thomas said. And, sometimes, scientists can tell if those arteries had hardened.
Simplified Text
The scans could help the team understand about ancient Egyptians' health. For example, some mummies still have arteries. **An artery is a tube that moves blood through the body.** **The artery could show if the person had been healthy or not.**

Table 3.1: An example of elaborative simplification containing two elaborations of varying contextual specificity. The sentence highlighted in green is an example of an elaboration with low contextual specificity, while the sentence highlighted in blue represents a highly contextualized elaboration, clarifying a statement in the original text.

define our labels and generate sequences at the sentence level.

Consider the example shown in Table 3.1. The original text snippet, taken from a news article, explains that scientists study mummy arteries to see whether they are hardened. In the corresponding simplified text snippet, we see two elaborations inserted – one, giving a simple definition of an artery, and the second clarifying the implication of hardened arteries. Effective elaborations must not only provide background, but they must do so in a contextual manner, adding relevant information to the surrounding text.

Often times, choosing when to elaborate is subjective; for example, in the same document about ancient Egyptian health from Table 3.1, the definition of an artery was inserted, whereas the concept of mummification was not. Though there is irregularity in *which* concepts are explained, there is a regularity in the manner in which concepts are elaborated.

We motivate our work by studying this regularity to understand *how* ideas, entities, or concepts are elaborated. At first glance, it seems that elaborative simplification might simply involve retrieving simple definitions or, even crafting informative post modifiers, as in Kang et al. (2019). However, effective and

informative elaborations take a variety of forms throughout the Newsela corpus involving reasoning over relevant document content. Not only are simple definitions of difficult concepts inserted, but often times, clarification or analysis sentences specific to document context are added in to aid comprehension or facilitate connections between content in the original text. The presence of these highly contextualized elaborations suggests that, in many cases, the retrieval and insertion of simple definitions is inadequate.

To address this spectrum, we construct a framework to categorize elaborations on a discrete scale ranging from low to high contextualization. We define contextualization as the degree to which a particular elaboration is specific to document context, drawing a distinction between contextual specificity and contextual relevance (as mentioned in the PoMo work (Kang et al., 2019)). All text that is in the simplified document is naturally contextually relevant, but will vary in how contextually *specific* the elaboration is. We use contextual specificity and level of contextualization interchangeably.

In our taxonomy, we have three levels of contextual specificity – elaborations with low contextual specificity tend to be definitions or standalone facts about entities or ideas in the original text. Non-definition details (i.e details added in to highlight impact, degree of severity) tend to be considered medium, and clarifications and analyses tend to be considered highly contextually specific. In our example about arteries from Table 3.1, the green sentence receives a rating of low, and the blue, a rating of high. Figure 3.1 contains more examples of elaborative simplification.

While inserting definitions may help provide background about entities, highly contextualized elaborations interpreting or clarifying content can help the reader understand the larger implications, connotations, or significance of ideas presented in original text. The example in Table 3.1 highlights two elaborations on opposite ends of the spectrum – the first elaboration requires little context, while

Original Text	Simplified Text	Contextual Specificity
Scottish voters rejected independence on Thursday, deciding to remain part of the United Kingdom after a historic referendum that shook the country to its core. The decision prevented a rupture of a 307-year union with England, bringing a huge sigh of relief to the British political establishment. Scots voted 55 percent to 45 percent against independence in a vote that saw an unprecedented turnout.	On Thursday, Scottish voters got the chance to decide whether to become independent. They voted to stay part of the United Kingdom. Scots voted 55 percent against independence. Forty-five percent voted in favor of breaking away from the United Kingdom. **The United Kingdom is a group of four countries. They are England, Scotland, Wales and Northern Ireland.** The portion of Scottish voters who showed up was greater than ever before.	Low
Those workers will not be paid until there's an agreement to fund the government anew. Unable to reach an agreement last night as the House and Senate played political tennis over a plan to temporarily fund the budget, the nation will wake up to an altered government landscape. Some of the services immediately affected are largely invisible, but important, nonetheless. The State Department, for example, will have to halt some processing of passport applications in federal offices not run by the agency but that are shut down, potentially threatening business or vacation travel of unsuspecting citizens.	At least 800,000 federal workers won't be able to work in the shutdown. At least temporarily. Those workers will not be paid until Congress agrees on a plan to pay for the government. The State Department, for example, will have to stop giving out some passports. **People need passports to travel to other countries.** This will disrupt their travel plans. The Treasury Department cracks down on crimes involving money. That work will have to take a break. But the Department of Transportation said that air traffic control services will continue.	Low
BEIJING " Investigators looking for Malaysia Airlines Flight 370 have put away their towed pinger locator and are about to call off searches for surface debris. Now, it's all up to a little yellow robotic submarine to find the missing Boeing 777 in an area bigger than the city of Los Angeles.	BEIJING " The search for the missing Malaysia Airlines plane is almost over. A yellow submarine will make the last try to find it. The plane disappeared a month ago. **It was traveling between the Asian countries of Malaysia and China.** Now, search parties are about to call off the searches on the ocean surface. Instead, they are sending a small submarine underwater.	Medium
WASHINGTON " The government shutdown has slowed or halted federal efforts to protect Americans' health and safety, from probes into the cause of transportation and workplace accidents to tracking foodborne illness. The latest example: an outbreak of salmonella in chicken that has sickened people in 18 states.	WASHINGTON " The government shutdown means many people are not working at some important federal agencies. Their absence makes it harder to protect Americans' health and safety. It is more difficult to look into the cause of accidents at work and on the roads and trains. **It is also harder to track diseases that come from food.** A case of salmonella in chicken is the latest example.	Medium
"As the delivery of health care evolves with an emphasis on better health outcomes, reducing chronic disease and controlling costs, CVS Caremark is playing an expanded role in providing care," Larry J. Merlo, the president and chief executive officer, said in a statement. "Put simply, the sale of tobacco products is inconsistent with our purpose." CVS, based in Woonsocket, R.I., also pledged to launch what it called a "robust national smoking cessation program".	The company has about 7,600 stores nationwide. It is the second largest drugstore company in the country. Only Walgreen Co. is larger. Larry Merlo is the president of CVS. **He said CVS wants keep people healthy.** Selling cigarettes goes against this goal. CVS also will start a national anti-smoking effort.	High
WASHINGTON " As a new school year begins, American parents should enthusiastically join first lady Michelle Obama's campaign for healthier school lunches " a campaign based on sound nutritional science with the goal of healthier, happier kids. The first lady has made improving childhood health through better eating and more exercise her signature issue. That's a wise choice, since childhood obesity reached epidemic proportions: In 2012, 1 in 3 American children were overweight or obese.	WASHINGTON " As a new school year begins, American parents should support first lady Michelle Obama. She has a new plan to make school lunches healthier. It is based on science and nutrition. The first lady wants to improve children's health through better eating and more exercise. **America has a weight problem.** Many people aren't just overweight, they are obese. Childhood obesity is a national problem: In 2012, 1 in 3 American children were overweight or obese.	High

Figure 3.1: Verified examples of elaborations of varying contextual specificity levels. Elaborations in the simplified text are in bold.

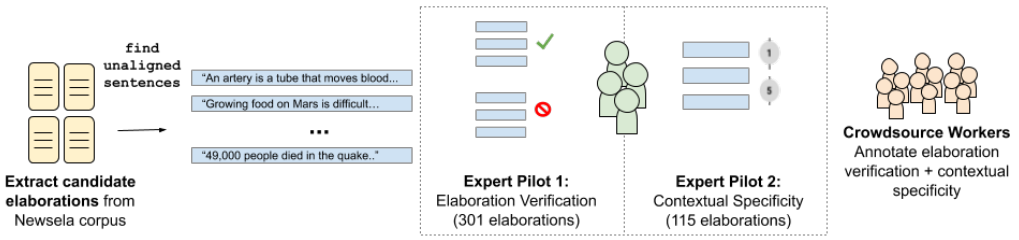


Figure 3.2: Corpus annotation flow. We first extract candidate elaborations from document sets in the Newsela corpus as described in Section 3.3.1. We then pass a subset of these candidates to experts for elaboration verification. We measure agreement between expert aggregate and ourselves and pass the rest of the candidates to crowdworkers to annotate for both elaboration verification and contextual specificity. In parallel, we ask a pair of experts to elaborate 115 elaborations in a contextual specificity pilot and measure agreement.

the second is highly contextualized, drawing a conclusion from content presented in the original text.

To this end, we propose two new modeling tasks related to elaborative simplification: contextual specificity prediction and elaboration generation. We explain task details and baseline modeling approaches in Chapter 4.

3.3 Data Resource Collection

In this section, we discuss construction of our dataset through automatic preprocessing, expert training and annotation, and crowdsourcing. We first describe strategies for trained annotators to extract elaboration sentences and annotate contextual specificity, and then discuss how we scale up the annotation of both using crowdsourcing with rigorous quality control. Figure 3.2 shows our full high-level annotation flow.

3.3.1 Extracting Elaborations

Detecting elaborative simplification requires crafting a way to reliably extract sentences containing new content in simplified documents. Ideally, human annotators

would read each sentence in the simplified document to assess whether sentence content is semantically new and missing from the original document. However, asking humans to read and annotate every sentence in each document in the Newsela corpus is prohibitively costly. To streamline this process, we first obtain candidate elaboration sentences with an automatic alignment heuristic, then ask human annotators to filter candidates to obtain true elaborations.

Candidate extraction. Each set of articles in the Newsela corpus consists of 4 or 5 simplified articles about the same event or topic written for varying grade levels, ranging from grade 3 to grade 12. This provides the opportunity for interesting studies of how elaborative simplification varies across different reading levels. However, to reduce the scope of this introduction of elaborative simplification, we leave this to future work, and choose the article written for the lowest grade level as our simplified document for every article set.

To identify candidate elaborations, we employ a heuristic based in our definition of elaborations (content that is *semantically* missing from the original document but present in the simplified document), using the approach from (Zhong et al., 2019). First, for each sentence in the simplified document, we find the sentence in the original document that is best aligned. That is, we calculate the similarity between each original-simplified sentence pair, and pick the sentence with the highest similarity score in a greedy manner. This increases the probability of detecting sentence splitting, when two or more adjacent sentences in the simplified document are aligned with the same sentence in the original document. Like Zhong et al. (2019), we use Sent2Vec (Pagliardini et al., 2017) to obtain sentence vectors for our cosine similarity computation, an unsupervised model for learning universal sentence embeddings from word vectors and character bigram vectors, trained on Wikipedia, tweets, and the Toronto book corpus. Zhong et al. (2019) use an alignment threshold of 0.94 when sentences are not split, and a threshold of 0.47

Simplified Text	Sent2Vec Region (Ours)	MassAlign Region
There may also be too many sea lions and not enough food. California sea lions used to be hunted for their fur and a special kind of fat called blubber. **The hunting nearly wiped them out.** In 1972, the government protected sea lions. Hunting them became against the law. The numbers of sea lions grew, Johnson said.	California sea lions were exploited in the 19th and early 20th centuries for their hides and blubber and continued to be hunted for sport in some areas later in the 20th century. The environment just may not be able to support a larger sea lion population. Although sea lions range from Mexico to Canada, the Channel Islands are where most of America's sea lions breed.	California sea lions were exploited in the 19th and early 20th centuries for their hides and blubber and continued to be hunted for sport in some areas later in the 20th century. The Marine Mammal Protection Act of 1972 led to dramatic increases in the populations of marine mammals, Johnson said. And the population is now abundant " 300,000, said NOAA Fisheries spokesman Jim Milbury, with a birth rate of about 50,000 a year.
China's Tiananmen Square changed forever during a few weeks in spring 1989. A popular leader named Hu Yaobang had just died. **Yaobang was a popular leader in China because he had changed many old rules and was making the government more modern**. Young people went to march on the square to show their sadness. More and more people showed up over the coming weeks.	The nature of the square changed forever during a few weeks in spring 1989. The death of former Communist Party General Secretary Hu Yaobang, a popular, open-minded reformer, led thousands of students and young people to march to the square in mourning. More and more protesters arrived over the coming weeks, eventually numbering hundreds of thousands.	In China, Maura Cunningham says, if you're going to hold an online discussion of the Tiananmen Square massacre, you better speak in code.
A new program in Chicago found jobs for some teenagers. Then a study showed that if teens have jobs, the number of violent crimes they do may go down. **Scientists wanted to see if having a job changes the way somebody acts.** So they picked teenagers. Then they gave each one a job for eight weeks.	The 730 teens who were offered jobs were picked at random from among 1,634 applicants. Their number of arrests for violent crimes was slightly lower than that of the remaining 904 teens while the jobs lasted, but the difference did not become statistically significant until six months into the study " three months after the jobs were completed. Heller said that the results may underestimate the impact of the jobs program, as a quarter of the teens who were offered jobs did not accept them, though some of those teens found other jobs on their own. "We explained that everybody's first job is a horrible job," Diaz said.	And not because the youths were too busy working to break the law. Those who were randomly chosen to get the eight-week positions were arrested for violent offenses 43 percent fewer times than their peers, and most of that difference occurred during the 13 months after the jobs were finished.

Figure 3.3: Comparison of alignment techniques – our alignment strategy using Sent2Vec (middle) tends to accurately capture aligned surrounding text, plainly exposing the elaboration, as opposed to that of MassAlign (Paetzold et al., 2017)
.

when they are. Their thresholds were calibrated using their manually aligned data. As their alignment strategy was used on the Newsela corpus as well, we use their thresholds and heuristics in our alignment step.

Through manual inspection, we found this Sent2Vec alignment strategy to be more accurate and reliable than that of other alignment algorithms, such as MassAlign Paetzold et al. (2017) and the Jaccard-based alignment algorithm proposed in (Xu et al., 2015). Figure 3.3 shows examples of the Sent2Vec-based alignment strategy compared to MassAlign. We hypothesize that MassAlign's drop in performance on our dataset could be attributed to the fact that the system

is heavily paragraph-based, only aligning sentences between detected aligned paragraphs.

We then consider sentences in the *simple* document that are *not* aligned with any sentence in the original document as **candidate elaborations**. Due to the fact that this system is largely based on the comparison of sentence embeddings, which may introduce noise (Arora et al., 2016), we find that this automatic system inherently has high recall.

Human Verification. In order to verify that the candidates proposed by our automatic alignment heuristic actually constitute elaborative simplification, we asked 13 native English speakers (henceforth called 'expert annotators') who are undergraduate students at our university to annotate 50 randomly selected documents (a total of 301 candidate elaborations) from our corpus. Each expert annotator annotated a subset of the 50 documents, and each candidate elaboration is annotated by 2 to 4 expert annotators. For each document set, the expert annotators were provided both the entirety of the original document, as well as the simplified document, and were asked to annotate each candidate elaboration as to whether they truly contain semantically new content. Our expert annotation interface is shown in Figure 3.4.

Alongside these annotators, we annotate 150 of these candidate elaborations ourselves. We measure agreement between the 13 expert annotators themselves using Krippendorff's alpha (Krippendorff, 2008), obtaining $\alpha = 0.36$. For the 150 sentences that we annotated ourselves, we aggregate the annotators' responses and measure agreement between their responses and ours using Cohen's kappa (Artstein and Poesio, 2008), obtaining $\kappa = 0.670$. While the agreement between annotators themselves is not very high, we find that the aggregated agreement is. This in line with other complicated tasks in NLP as described in Nye et al. (2018). Per the high aggregated agreement, we use human aggregate labels for each example in

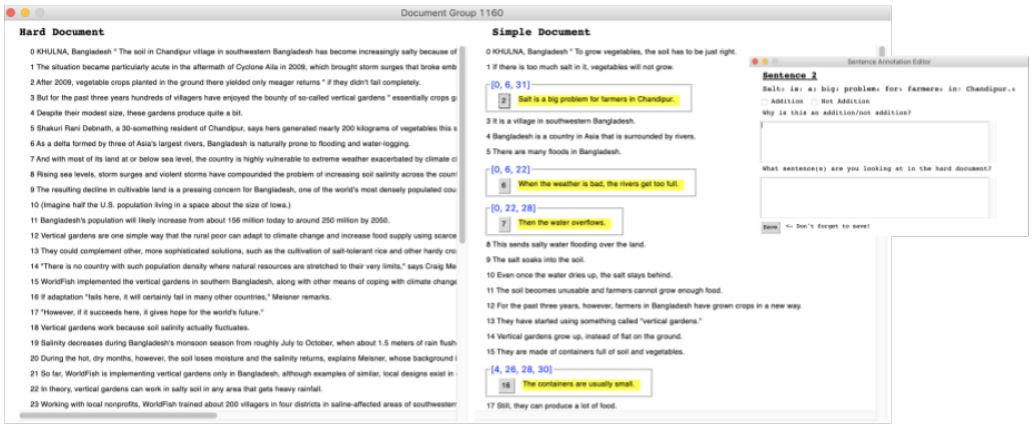


Figure 3.4: Expert annotation interface for semantic addition/not addition. For each document set that expert annotators were asked to annotate, we provide them the entirety of the original and simplified documents. We highlight candidate elaborations in yellow and suggested sentence regions in the original document.

our dataset.

3.3.2 Contextual Specificity

To understand the varying levels of contextualization, we ask a pair of experts from the previous pilot to annotate 115 randomly selected verified elaborations. As mentioned, we define contextualization as the degree to which a given elaboration is specific to the context. Each expert was given, again, the entirety of the original and simplified documents as well as the highlighted elaboration, and asked to label the contextual specificity of the elaboration on a scale of 1 to 5. We measure agreement between this pair of experts using Spearman's correlation, obtaining $\rho = 0.72$. Their aggregate labels are published as a part of our test set. We include the full instructions we provide to experts and crowdworkers in the appendix.

3.3.3 Crowdsource Annotation

Elaboration verification and contextual specificity rating are difficult annotation tasks, both needing careful reading and thoughtful reasoning over text. For both

the pilots described in the previous sections, we provided training in the form of thorough instructions, example documents and annotations, and a few practice sentences. While these trained annotators provide high quality reliable annotations, they ultimately cannot annotate a dataset of the scale supervised learning systems require. To remedy this, we use crowdsourcing platform Amazon Mechanical Turk [1] to collect examples at scale. In this section, we discuss task design and strategies for quality control.

Task Setup At a high level, for each candidate elaboration, we provide crowdworkers with 5 to 7 sentences containing the elaboration from the simplified document, and the corresponding snippet of text from the original document. We ask workers to annotate both elaboration verification and contextual specificity in a single HIT. Figure 3.5 contains a snapshot of the annotation interface we provide crowdworkers.

Incorporating feedback from our expert pilots, we determined that providing the whole document was often distracting, proving necessary only in rare cases where content was drastically rearranged. We instead choose to display a corresponding region from the original document as opposed to the whole text. To select the region to display, we apply the following heuristic: we identify the window of 5 sentences before and after the candidate elaboration, referred to as the "simplified region". For each sentence in the simplified region, we obtain the corresponding aligned sentence in the original document. Because these sentences may not always be contiguous, we partition the alignment into "clusters" of sentences, using a manually calibrated testing window threshold of 5 sentences. For example, two clusters would be examined for the aligned sentence set in the original document of $(1, 4, 15, 17, 19)$ – a cluster containing $(1, 4)$ and $(15, 17, 19)$. We then select the larger cluster, filling in gaps, and padding with two sentences

[1] https://www.mturk.com/mturk/welcome

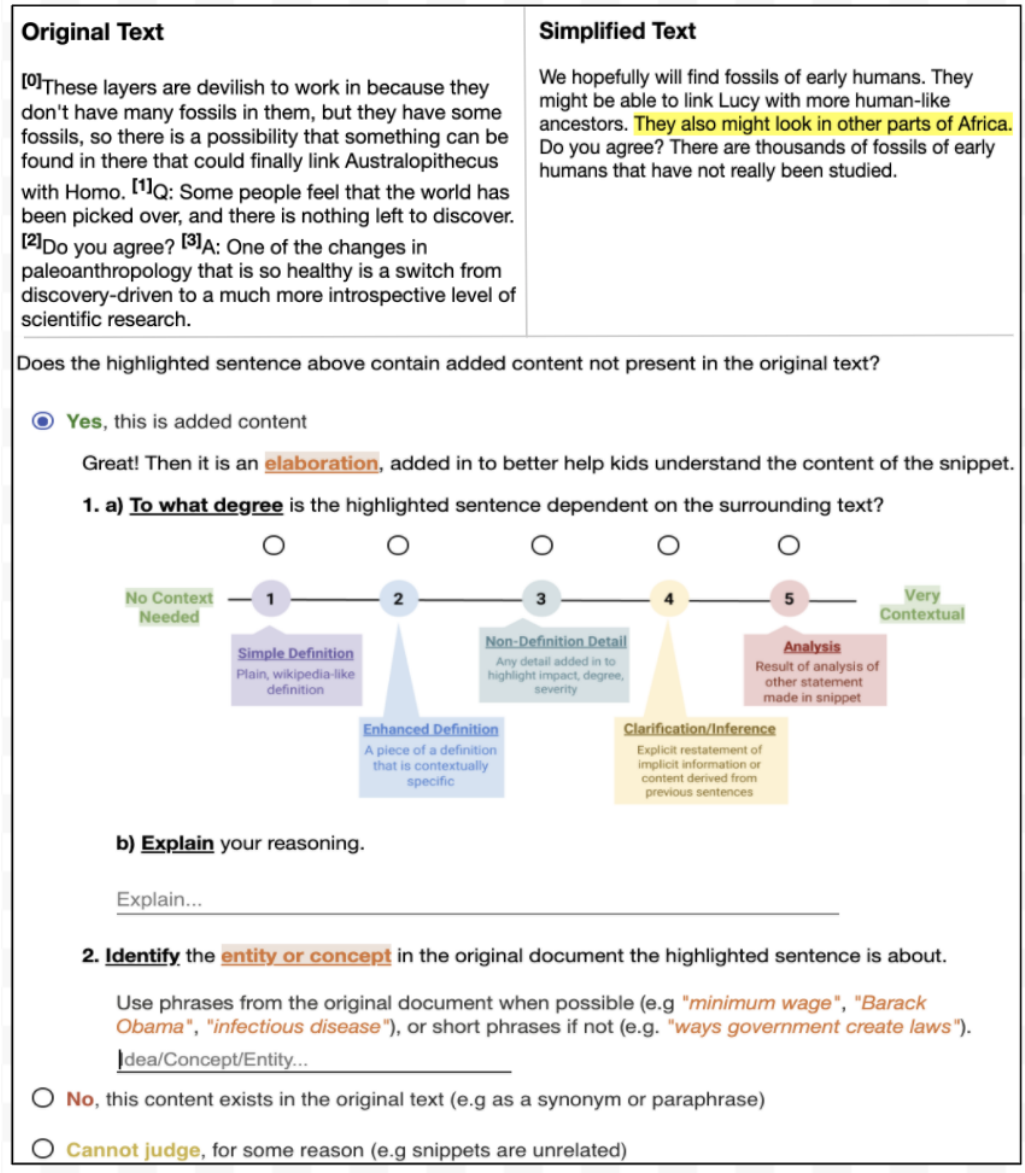


Figure 3.5: Annotation Interface for crowdworkers. We display a fine-grained rating scale of 1 - 5, but for the purposes of analysis and modeling, we bucket a rating of 1 and 2 as low, 3 as medium, and 4 and 5 as high contextual specificity.

before the first aligned sentence and the last aligned sentence in the region.

Crowdworkers were asked to categorize each candidate as a true elaboration, not an elaboration, or indicate that the snippets were unrelated. Upon indicating that a candidate was indeed an elaboration, we asked them to rate the contextual specificity of the elaboration according to the scale in Figure 3.5, and enter the entity from the original text that was being elaborated. To streamline entity entry, we provide crowdworkers with an auto-complete system populated with noun chunks from the original text snippet, similar to FitzGerald et al. (2018).

Quality Control As discussed, elaboration verification and contextual specificity rating require careful reading and reasoning. To ensure high quality annotations, we ask crowdworkers to provide a rationale for each rating decision, as proposed by McDonnell et al. (2016). These rationales provide insight into worker interpretations of our task, allowing us to actively curate annotations to only include reliable annotations in our dataset. For example, using this method, we were able to remove annotations where crowdworkers inflated contextualization ratings due to coreferent mentions of entities (i.e *"It is a tube that moves blood"* as opposed to *"An artery is a tube that moves blood")*.

In addition, we require all crowdworkers to reside in the United States, Great Britain, Canada, Australia, or New Zealand, and to have completed greater than 100 HITs with an acceptance rate of 95%. We have each elaboration annotated by 5 unique crowdworkers.

Label Collapsing. The annotation interface we provide to experts and crowdworkers consists of a contextual specificity rating scale ranging from 1 to 5. After collecting all of our data, we observed that the crowdworkers found boundaries between adjacent ratings slightly blurry. This is reflected in Figure 3.6, obtained by calculating raw disagreement rates between crowdworkers for each pair of possible

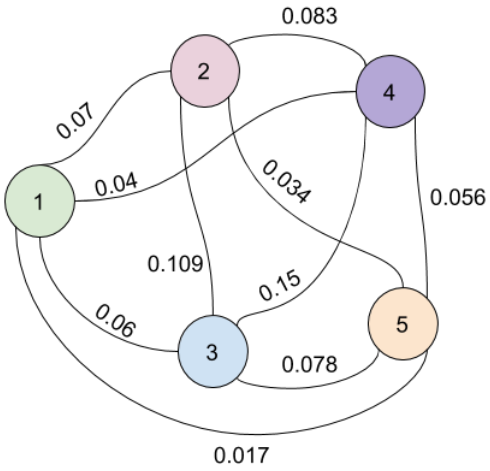


Figure 3.6: Disagreement rates between crowdworkers for each pair of possible contextual specificity ratings as a proportion of total annotation pairs. Agreement rates (i.e self-loops in the graph) are not shown.

contextual specificity ratings as a proportion of total crowdworker annotation pairs per candidate elaboration. To address this, and decrease modeling complexity, we collapse our 1 - 5 rating scale to 3 labels, binning 1 and 2 together, and 4 and 5 together. This binning scheme naturally reflects low, medium, and high contextualization, and reduces total raw disagreement by 13%. We experimented with several other binning schemes as shown in Table 3.2, and found that this scheme had the highest correlation with a subset of 112 expert annotated contextual specificity ratings. For modeling and analysis, we utilize these collapsed labels.

Agreement between trained and crowdsource annotators. To aggregate crowdworker labels for contextual specificity, we first collapse labels into low, medium, and high contextual specificity according to the scheme in the previous paragraph. We use Krippendorff's alpha with an ordinal distance metric (Krippendorff, 2008) to measure agreement between crowdworkers and experts, aggregating Turker responses and expert responses to obtain a value of $\alpha = 0.47$, indicating moderate agreement (Artstein and Poesio, 2008).

Binning Scheme	Agreement (α)	Corr (ρ)
1, [2-3], 4, 5	33.85	58.27
[1-2], 3, [4-5]	**49.84**	**70.0**
[1-2], 3, 4, 5	43.29	67.56
1, [2-3], [4-5]	47.22	59.27
[1-2], [3-4], 5	48.60	67.02

Table 3.2: Agreement statistics between crowdsource labels binned according to various binning schemes and expert labels on 112 randomly selected, verified elaborations. We measure agreement using Krippendorff's alpha (Krippendorff, 2008) and Spearmean's correlation.

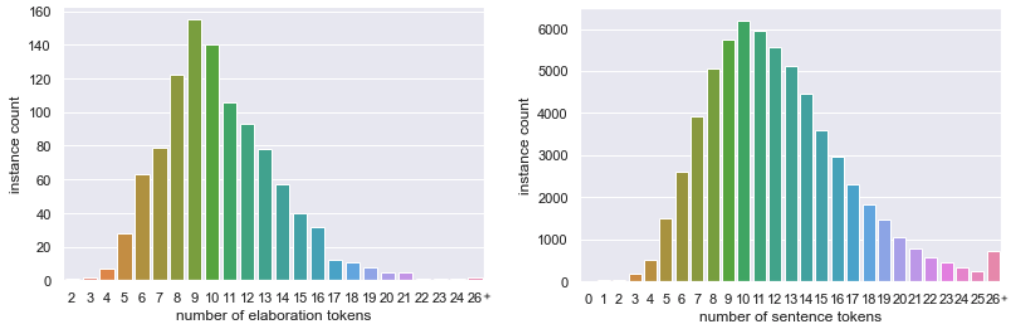
We attribute the disparity between inter-expert agreement and expert versus crowdsource aggregate agreement to be due to subjectivity of this task, especially amongst untrained crowdworkers. Though crowdsourcing our data does result in a slightly noisier training set, we are able to collect data for supervised learning and analysis at scale.

3.3.4 Dataset Statistics

Using Mechanical Turk, we annotated over 4K candidate elaborations, establishing an approximate 30% conversion rate from candidate elaborations to verified elaborations. We collected around 1300 true elaborations, and randomly split them into train, validation, and test sets. Table 3.3 shows our final train, validation, and test set distributions over different levels of contextual specificity. Our validation and test sets both contain expert labels for contextual specificity.

	Low	Medium	High	**Total**
Train	303	349	397	**1049**
Valid	71	39	24	**134**
Test	32	35	49	**116**
Total	406	423	470	**1299**

Table 3.3: Dataset distribution by contextualization level for train, validation, and test splits.



a Elaboration Token Length Distribution b Global Sentence Token Length Distribution

Figure 3.7: Length distributions of elaborations vs non-elaborations in the simplified Newsela corpus.

We find that on average, elaborations consist of 10 tokens, compared to an average length of 12 tokens for all sentences across the set of simplified documents (see Figure 3.7). This again indicates the diversity of elaborations.

3.4 Elaboration Qualitative Analysis

In this section, we present a series of analyses to study the nature of elaborations, including their varying levels of contextual specificity, the type of knowledge encoded in elaborative sentences, as well as the entity, idea, or concept being elaborated.

3.4.1 Contextualization

Studying contextualization can aid elaboration generation systems in numerous ways, such as developing a reranking system to generate text based on the type of elaboration desired, incorporating a retrieval module for factual and detail-based elaborations, integrating multi-hop reasoning modules for highly contextualized explanations, or to train contextualization embeddings to condition on during generation.

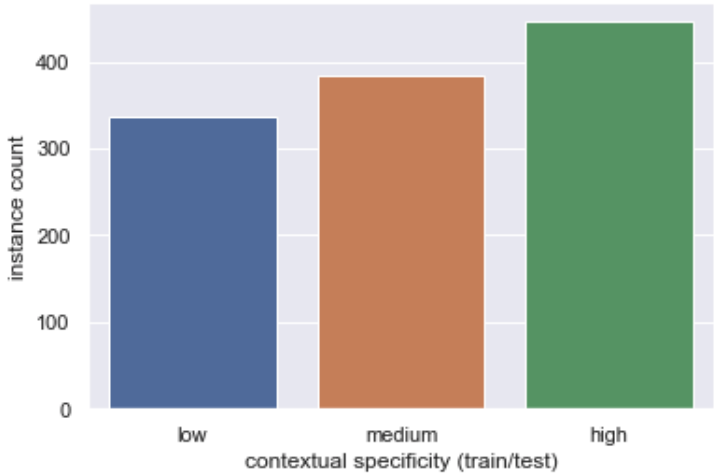


Figure 3.8: Contextual specificity distribution across train and test splits. We can see that the majority of elaborations are highly contextualized, requiring some form of reasoning over content in the original document.

As shown in Figure 3.8, the majority of elaborations in our corpus are highly contextualized, indicating that effective simplification requires a nuanced understanding of text in the original document. Highly contextualized elaborations occur most frequently in our dataset in the form of clarifications and analyses of content from the original text, affirming the fact that a definition generation system is not adequate for elaborative simplification.

3.4.2 Knowledge Type

While studying contextualization provides an effective framework for understanding *how* things are elaborated, it does not provide insight into the type of content embedded in elaborations (i.e commonsense knowledge, scientific knowledge). To further understand the nature of elaborative simplification, we conduct a small study on a subset of our data to understand the *type* of knowledge in elaborative sentences. Studying the type of knowledge inserted during elaborative simplification can help further inform the type of systems necessary for elaboration generation.

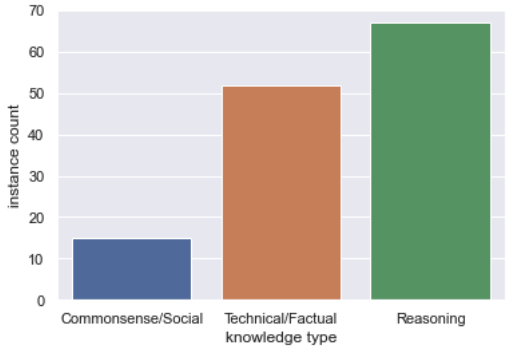


Figure 3.9: Distribution of knowledge type aggregate crowdworker labels across 134 randomly sampled verified elaborations. We see that reasoning using knowledge present in the original document is most prevalent.

We randomly sample 134 verified elaborations from our dataset and ask crowdsource workers to label them according to a simple taxonomy we manually crafted consisting of three categories. Examples of these categories can be seen in Figure 3.10.

1. **Technical/Factual Knowledge:** Content which children may not have encountered in school, textbooks, or the internet. This category includes mostly factual or scientific information.

2. **Commonsense/Social Knowledge:** Content containing knowledge obtained from living in society including social norms, expected emotions, or ordinary encounters adults must deal with.

3. **Reasoning using knowledge in the original document:** Content added to simply aid comprehension of the original document in the form of explicit inference. No specific outside knowledge is added to the document.

We collect these annotations in an identical manner to our main elaboration annotations – we have 5 workers annotate each example, and apply the same restrictions as stated in Section 3.3.3. Figure 3.9 shows the result of this study. The

majority of elaborations contain technical/factual knowledge as well as reasoning statements. While knowledge type may be subjective and more challenging than contextual specificity to use as a framework for analyzing elaborative simplification, it does provide more insight into the kind of content being added into these documents during the simplification process.

To investigate the relationship between contextual specificity and knowledge type, we calculate Spearman's correlation between the number of workers that labeled an example a particular knowledge type and the gold contextual specificity level. Between scientific/factual knowledge and contextual specificity there is a $\rho = -0.42$ correlation, between reasoning and contextual specificity a $\rho = 0.27$ correlation, and between commonsense/social reasoning, a $\rho = 0.17$ correlation. This suggests that scientific/factual knowledge does not tend to be highly contextualized (per the inverse relationship), while commonsense reasoning and inference tend to be contextually specific.

3.5 Summary

In this chapter, we presented elaborative simplification, a phenomenon which we observe as a part of document simplification. To aid our study, we described our process of collecting a corpus of over 1.3K elaborations through a combination of automatic processing and human verification, and performed a qualitative analysis of contextual specificity and type of embedded knowledge.

Original Text	Simplified Text	Knowledge Type
LOS ANGELES " California and New York still lure hundreds of thousands of immigrants from across the globe, but Texas, Florida, Colorado and the Carolinas are far more magnetic for people already living in the country, according to new estimates released Thursday by the U.S. Census Bureau. Late last year, the Census Bureau announced that Southern and Western states had driven much of the population growth nationwide.	But the U.S. Census Bureau says that Americans find Texas, Florida, Colorado and the Carolinas far more appealing. The Census Bureau counts the number of people who live in the United States. **It also tracks where they live and move to.** Its latest report found that states in the Southern and Western U.S. helped the country's population grow the most. More babies were born in those states than in the North and East.	Technical/Factual Knowledge
Patriots Samuel Adams and Paul Revere took part in the original ceremony, when a cowhide capsule was placed as the state moved from its old statehouse to its new one across from the Boston Common. The contents are thought to include a collection of coins dating from 1652 to 1855, when the cowhide capsule was replaced with a metal box.	America's famous patriots Samuel Adams and Paul Revere were there. Paul Revere made a famous horseback ride in 1775. **He warned people that the British were coming to attack.** "It's exciting," Bill Galvin said. He is a Massachusetts official.	Technical/Factual Knowledge
The researchers even tried out their rover on elephant seals, who didn't budge when a rover came close to their heads or tails (which is where they are usually tagged). That's a good sign; as a rule, an elephant seal does not react kindly to someone approaching its backside. Such robots could be used to investigate the lives of all kinds of animals without disturbing them the way a human scientist's presence would, the study authors wrote.	On a recent day, they used the robot with some elephant seals. The seals did not seem to notice the robot coming near their heads or tails. **Scientists were pleased.** Elephant seals usually do not like it when someone or something gets close to their backsides. It is exciting to think about how robots can help humans learn more about animals in the wild.	Commonsense/ Social Knowledge
Duncan did not have a fever when he left Liberia on Sept. 19, but developed symptoms days after arriving in Dallas. He first sought medical care the night of Sept. 25, but was sent home with antibiotics. When his condition worsened on Sept. 28, he was rushed back to Texas Health Presbyterian Hospital Dallas, where he was in isolation in critical condition. He had been receiving an experimental treatment using the antiviral drug brincidofovir.	Three days later, he became sicker and was rushed back to Texas Health Presbyterian Hospital Dallas. He was in a room by himself in the hospital. **He must be kept away from the other patients because the disease could spread.** Duncan was extremely ill. Because doctors did not realize Duncan had Ebola, many are afraid.	Commonsense/ Social Knowledge
Sunday's quake, which erupted 50 miles off the coast, caused light to moderate shaking. No injuries or damages were reported. As for the next 9.0 quake, US Geological Survey seismologist David Oppenheimer said: "It could be today. It could be 100 years from now".	On March 16th, a small earthquake caused by the Cascadia fault erupted 50 miles off the coast, causing light shaking. No injuries or damages were reported. **The West Coast was spared this time.** Scientists still wonder when "The Big One" will hit.	Reasoning using knowledge in the original text
In Antarctica, the scientists studied the reactions of king penguins on Possession Island. A human who invaded a penguin's personal space caused the bird's heart rate to spike much higher than a rover did, the researchers found " and the effect from the human encounter lasted much longer. "Human approaches led to an excess in (heart rate) approximately four times larger than that due to rover approaches," they wrote.	Researchers are using a robot with groups of king penguins in Antarctica. They are watching to see what penguins do when a robot gets close. **Then they compare it with how the penguin acts around humans.** They found that penguins are more scared of humans than robots. When a king penguin feels scared, it tries to run away.	Reasoning using knowledge in the original text

Figure 3.10: Knowledge type examples annotated by crowdworkers.

Chapter 4

Modeling and Experiments

In this chapter, we introduce two tasks related to elaborative simplification: contextual specificity prediction, and elaboration generation. We discuss baseline modeling and experiments for both tasks, and reflect on the ability of large-scale pre-trained language models (LMs) to generate a range of contextually specific elaborations.

4.1 Contextual Specificity Prediction

One essential component of generating effective elaborations is to produce explanations of varying contextualization, as introduced in 3.4.1. Predicting contextual specificity can help downstream when selecting from a set of candidate generations, or to train contextualization embeddings to condition on during generation. In this section, we explore methods for contextual specificity prediction.

To understand the role of context, we first explore input without the elaboration sentence. This resembles a practical scenario, where elaboration sentences are generated during the process of text simplification. We later incorporate the actual elaboration in tandem with document context for downstream application, for example to re-rank candidate elaborations.

4.1.1 Methods

Contextualization prediction involves a classification task[1] – given a snippet of text around and including the elaboration, the model predicts the elaboration's level of contextual specificity. For each elaboration, our contextual specificity label $l \in \{\text{low}, \text{medium}, \text{high}\}$. We leverage large scale pre-trained language models and explore two different settings for predicting level of contextual specificity, one setting based only off of context, and the other based on surrounding text *and* the actual elaboration.

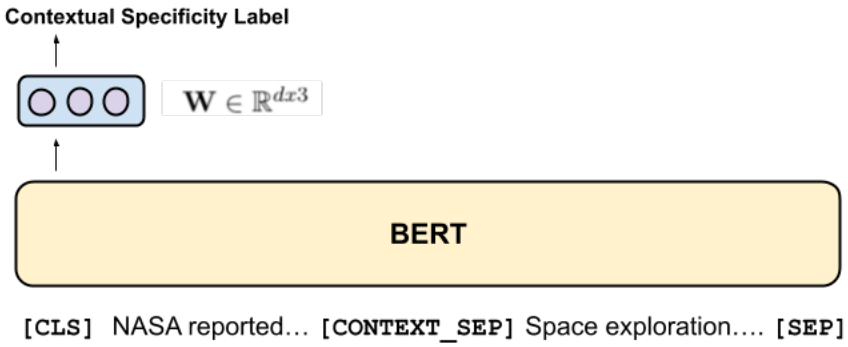


Figure 4.1: Contextual specificity classification model using BERT's `[CLS]` token. Example input shown is context-based (elaboration free). We train a special context separator token to separate between context from the original and simplified documents.

Without Elaboration To understand the role of text surrounding the elaboration in predicting contextualization, we first build a classification model using context from both the original and simplified documents as input. While determining contextual specificity clearly involves utilizing the elaboration itself, we explore excluding it from model input in this setting to understand whether contextualization is a predictable phenomenon from context alone. This can be

[1] We have experimented with ordinal regression methods but classification yielded much better results.

seen as a largely realistic setting, when during simplification, only prior text is available to the model, before the actual elaboration text is generated.

We vary the number of sentences we use prior to the elaboration in the simplified document, and experiment with 3 different ways of passing input to the classifier:

1. **CLS token of original and simplified snippets:** Language models we use for this task utilize a special `[CLS]` token for classification tasks. We use this token, along with learning a representation for a separation token `[CONTEXT_SEP]` to distinguish between the two spans of context. We feed in input to the model as follows: `[CLS]` a `[CONTEXT_SEP]` b, where a corresponds to the sentences in the region aligned in the original document, and b corresponds to the sentences leading up to the elaboration in the simple document.

2. **CLS token of simplified snippet** We experiment with feeding sentences only from the simplified document as context to the model to predict contextual specificity. We do this in two settings – feeding 2 sentences and 4 sentences as input to the model respectively.

3. **CLS token of original snippet** In addition to experimenting with only the simple document, we explore feeding only the aligned region from the original document as input to the model. We trim the aligned region such that it ranges from 4 - 7 sentences at maximum.

With Elaboration We also build a contextual specificity model including elaborations to utilize downstream in other related tasks, such as reranking generated elaborations. We use a combination of text from the simplified document and the original document in addition to the elaboration sentence itself as input to our model.

1. **CLS token of only the elaboration:** We experiment feeding only the `[CLS]` token of the elaboration sentence as input to the model to understand how much the model relies on the elaboration itself to predict contextualization.

2. **CLS token of original/simplified snippets and elaboration:** Language models we use utilize the `[SEP]` token to discern between distinct sequences of text. In this setting, we feed in input to the model as follows: `[CLS]` a `[CONTEXT_SEP]` b `[SEP]` c, where a corresponds to the sentences in the region aligned in the original document, b corresponds to the sentences leading up to the elaboration in the simple document, and c is the elaboration itself.

3. **CLS token of simplified snippet and elaboration:** To understand whether contextual specificity depends on the original document at all, we remove a from the previous setting.

4. **CLS token of original snippet and elaboration:** We experiment with feeding in only the snippet from the original document and the elaboration itself in the same form as the previous setting.

4.1.2 Experiments

We utilize the base version of BERT (Devlin et al., 2018) via the HuggingFace Transformers library (Wolf et al., 2019) to model contextual specificity. We train our baselines on the train split of our dataset, and feed the sequence representation from the `[CLS]` token embedding into an output layer for classification[2]. After tuning on the validation set, we train our model (4.1) in each setting for 3 epochs, using a batch size of 32 and a learning rate of 1e-3. We use the default dropout

[2]We tried finetuning our contextual specificity prediction models on our elaboration dataset, but found that our dataset was too small to yield stable results.

		Acc.	F1	Correlation	MAE
Context Only	Original + Simple	41.6	40.4	25.8	0.774
	Simple	44.9	44.0	29.6	0.705
	Original	35.3	29.7	-	0.895
With Elaboration	Elaboration	50.4	47.6	35.9	0.630
	Original + Simple + Elab	49.5	46.8	35.7	0.670
	Simple + Elaboration	**53.6**	**52.1**	**46.9**	**0.595**
	Original + Elaboration	43.0	39.3	23.3	0.752

Table 4.1: Contextual Specificity Prediction results, including accuracy, F1, Spearman's correlation, and Mean Absolute Error. We bold our best results.

rate of 0.1 for self-attention layers, but refrain from adding dropout on our linear layer. We evaluate model performance using F1, correlation, and mean absolute error, and report mean performance over 15 different, randomly initialized runs.

4.1.3 Results

Results across all settings are shown in Table 4.1. We see that the best predictor of contextual specificity is context in the form of 4 sentences before the elaboration, combined with the elaboration itself.

Amount of simplified text. For each context setting involving sentences from the simplified document, we vary the number of sentences (s = 4, and s = 2) before the elaboration, and report results from the better of the two. In almost all settings, adding in a larger window of sentences from the simplified document boosted results, suggesting that more context is better. Through experimentation, we find that decreasing the amount of context from the simplified document results in a drop in performance.

Elaboration presence. Our results show that performance drops without the elaboration itself, suggesting that the language used in the elaboration itself is highly indicative of contextualization. This indicates that contextual specificity

prediction can be helpful during generation, as the language used in the generated elaboration itself plays a large part in dictating contextualization. As expected, the sentences from the original document are a poor predictor of contextualization, suggesting that content itself is not necessarily a strong indicator of contextual specificity. The best predictor of contextualization based on context alone seems to be 4 sentences from the simplified document prior to the elaboration. In a practical setting, elaborations of varying contextualization could be suggested based on the window of previous sentences written during simplification.

Original text presence. In all settings in which the aligned snippet of text from the original document was fed in as partial or complete input to the model, we see a reduction in performance. Compared to text from the simplified document, text from the original document is stylistically distinct. Consequently, when jointly fed in as context with text from the simplified document, the input is largely incoherent. We leave resolving this in a natural way as future work.

4.2 Elaboration Generation

We now move to the task of elaboration generation. Much of recent work on pre-trained language models has shifted focus to understanding the world knowledge and commonsense that these systems capture (Shwartz et al., 2020; Davison et al., 2019). Additionally, it is still unclear if these systems are able to perform *multiple* steps of reasoning, as some highly contextualized elaborations in our dataset require. We investigate the abilities of large-scale pre-trained LMs to address some of the challenging aspects of elaboration generation, including their ability to produce a range of contextually specific elaborations, perform commonsense reasoning, and generate relevant and coherent content given snippets of context from documents in our corpus.

4.2.1 Methods

Elaboration generation ultimately involves generating a sequence conveying some useful new content, conditioned on some variable-length document context. We explore the ability of large scale pre-trained language models (LMs) to generate useful, linguistically coherent, and semantically plausible elaborations. In addition, we investigate the contextual specificity of generated elaborations, and incorporate our contextualization prediction model from the previous section during decoding.

Finetuning. We explore LM performance across two settings. First, we organically elicit elaborations in a zero shot setting, without fine-tuning on our dataset. Then, we finetune the LM on the set of simplified documents in the Newsela corpus, as well as on our dataset of verified elaborations [3]

Context. To understand the role that document context plays in elaboration generation, we elicit elaborations from the language model by providing it:

1. **s2**: Two sentences prior to the gold elaboration in the simplified document
2. **s2_h**: Concatenation of two sentences prior to the gold elaboration in the simplified document, and the corresponding aligned region in the original document
3. **s4**: Four sentences prior to the gold elaboration in the simplified document

Contextual Specificity Reranking. To investigate the importance of contextual specificity in generating effective elaborations, we incorporate our best contextual specificity prediction model from 4.1. For each example in our test set, we first generate elaborations in a greedy manner, and compare that to a

[3] We performed experiments with finetuning on our dataset alone, but found that the performance dropped in comparison to finetuning only on the Newsela simplified corpus.

re-ranking setting in which we generate sequences with beam search, and pick the highest likelihood sequence that matches the gold contextual specificity level. In practice, one would ideally use a contextual specificity model trained *without* the elaboration itself. However, since we leave to future work to build a strong model presented with this setup, we instead explore the upper bound with the generation experiments.

Evaluation. We use BLEU score (Papineni et al., 2002) as our automatic evaluation metric for elaborations generated by the LM. While BLEU score does provide a quantitative metric to capture overlap between the generated and gold elaboration, it fails to capture semantic similarity – contextually specific elaborations that provide useful and new content, but that don't necessarily overlap with the gold elaboration are penalized. To remedy this, we perform a human evaluation study with our best performing model setting. We provide a pair of evaluators with two sequences (greedily decoded, and the most likely sequence in the beam matching the gold contextualization level), and ask them to select the sequence they thought was the most coherent, relevant, semantically plausible, and elaboration-like. We allow multiple selection if both sequences are equally good, and no selection at all if both sequences are poor, following a human evaluation setup similar to Panthaplackel et al. (2020).

4.2.2 Experiments

We use GPT-2 medium (Radford et al., 2019) from the HuggingFace Transformers library to finetune and generate elaborations. For fine-tuned settings, we finetune for 3 epochs with a learning rate of 1e-5 and a batch size of 32. We experiment with two different decoding strategies - greedy, and beam search with a beam size $b = 5$ [4] Results are shown in Tables 4.2 and 4.3. We report corpus BLEU-1 and

[4]We did conduct experiments with increased beam sizes, but found little gain, as for the purposes of our demonstration with contextual specificity, we selected the highest likelihood

BLEU-2 scores on our test set.

Human Evaluation Our human evaluation metric is reported as the percentage (out of 116, our test set size) for which humans chose the sequence as higher quality. These sequences were generated using our best performing model, GPT-2 medium finetuned on the Newsela simplified corpus, with two sentences before the gold elaboration as context. The results of this study can be seen in Table 4.3. We calculate human agreement via Cohen's kappa with MASI distance (Passonneau, 2006), obtaining a value of $\kappa = 0.41$, indicating moderate agreement (Artstein and Poesio, 2008).

No Finetuning				
	Greedy		Reranking *(b=5)*	
	BLEU - 1	BLEU - 2	BLEU - 1	BLEU - 2
s2	12.48	2.71	13.64	3.71
s2_h	12.21	2.59	11.43	2.93
s4	13.46	3.35	14.48	4.12

Table 4.2: BLEU-1 and BLEU-2 scores for GPT2-Medium, without finetuning.

Simple Corpus Finetuning						
	Greedy			Reranking *(b=5)*		
	BLEU - 1	BLEU - 2	H (%)	BLEU - 1	BLEU - 2	H (%)
s2	20.79	6.77	49.0	**20.98**	**7.30**	**57.27**
s2_h	18.68	5.66	-	16.74	5.20	-
s4	20.82	5.54	-	20.12	6.92	-

Table 4.3: BLEU-1 and BLEU-2 scores for GPT2-Medium, with finetuning on the set of simplified documents in the Newsela corpus. Results for our best model, which we conduct human evaluation on, are in bold.

sequences with a predicted contextual specificity level matching the gold.

4.2.3 Discussion

Mirroring contextual specificity prediction, we observe that our best elaboration generation model involves context from the simplified document only, this time with only two sentences provided. We attribute the drop in performance between models with and without the original text as a part of input largely to the crude incorporation of content from the original document, which is stylistically starkly different from simplified text, most notably in terms of length and vocabulary complexity. We leave including content from the original document in a natural way during generation as future work.

Qualitative assessment of generated sequences. After finetuning on the Newsela simplified document corpus, we observe that GPT-2 is able to adopt elaborative style (i.e short sentences of 7-13 tokens long, limited vocabulary, etc), as expected. In addition, we find that the model can be effective at generating simple definitions when it correctly identifies an entity in the provided context. While the style of generated elaborations does match the simplified corpus, the content is often not anchored in the document itself; generated sequences are ostensibly relevant to the snippet of context provided, but seem irrelevant when placed in the larger document. While our reranking method does take contextual *specificity* into account, it does not take *factuality* or *relevance* into account. An improved reranking scheme considering both of these facets could promote sequences that better align with the larger document context.

Reranking Improvement. For most of our models, we do see an improvement when sequences are re-ranked according to contextual specificity. Our human evaluation study between greedily generated sequences and re-ranked sequences after finetuning GPT-2 shows this increase in performance as well. This suggests the importance of contextual specificity in improving elaboration generation, and

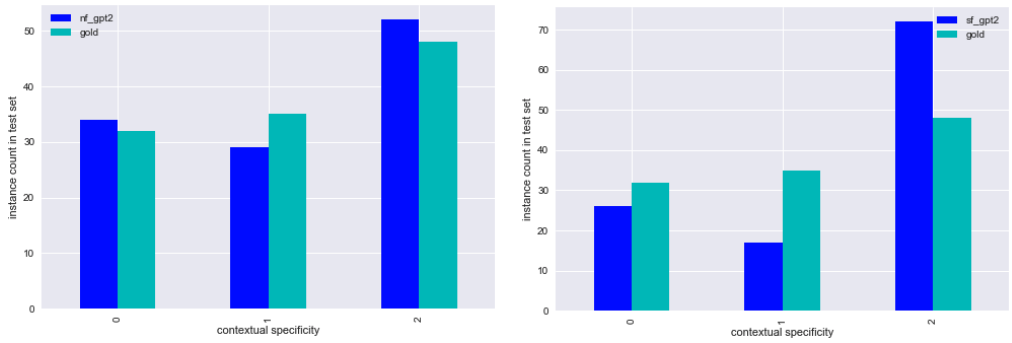
highlights the need for further improvement of contextualization modeling. Examples of this can be seen in Figure 4.2. In order to make reranking truly useful, we need to increase the reliability of contextual specificity prediction given an input combination of document context and the candidate elaboration.

Original Text	Simplified Text	Generated Text (s2)
Diplomatic posts in five other countries would reopen Monday, the State Department said, including those in Afghanistan and Iraq, where terrorist attacks have been frequent. State Department spokeswoman Jen Psaki said the extended closings were 'not an indication of a new threat stream.' Given that a number of our embassies and consulates were going to be closed in accordance with local custom and practice for the bulk of the week for the Eid celebration at the end of Ramadan, and out of an abundance of caution, we've decided to extend the closure of several embassies and consulates," she said.	The United States closed embassies in several Muslim countries over the weekend. The government made the decision after getting information about a possible attack. 'The U.S. also warned travelers about the threat. The embassies are where the official U.S. representative to a country works. **The grounds of a U.S. embassy are considered U.S. soil.** The closings are only for a few days.	**Gold:** The grounds of a U.S. embassy are considered U.S. soil. **Greedy:** `The U.S. government has a lot of information about the people who live there.` **Reranking:** `The U.S. Embassy is in Washington, D.C.`
It's heartbreaking that it's come to this, but it's the right thing to do," Paul Gibbons, managing director of the nonprofit Turtle Conservancy's Behler Chelonian Center in Ventura County, said as he gently placed a 30-pound adult female ploughshare tortoise on a small table [...] The tortoise was branded for life, which in her case would be roughly 160 years. We've blemished her natural beauty, so she's just a number in a system now," Gibbons said.	A tortoise can live 160 years. It's heartbreaking that it's come to this. But it's the right thing to do," Gibbons said. He is in charge of the Turtle Conservancy in California. **A conservancy saves animals from dying out.** We've spoiled her natural beauty," Gibbons said.	**Gold:** A conservancy saves animals from dying out. **Greedy:** `It's the right thing to do.` **Reranking:** `The conservancy is in charge of protecting the world's largest turtle.`
Putin once again called the change of power in Kiev last month, when Ukrainian President Viktor Yanukovych fled after the bloody crackdown against months-old protests had left scores of people dead, a coup d'etat and said that the interim government of Ukraine was illegitimate and not worth talking to. 'He blamed the U.S. and Western Europe, saying they encouraged coup plotters and had "crossed the line in the case of Ukraine. They acted roughly, irresponsibly and unprofessionally," Putin said. They knew quite well millions of Russians live in Ukraine and Crimea.	He blamed Ukraine's new leaders for the takeover. He called them names -- "nationalists, Russophobes and anti-Semites." A Ukrainian nationalist is a patriot. A Russophobe is anti-Russian. **An anti-Semite is someone who hates Jews.** In February, Ukrainian President Viktor Yanukovych fled Ukraine. He left after angry protests against his government. Many people were killed. Late last month, Russian soldiers marched into Crimea.	**Gold:** An anti-Semite is someone who hates Jews. **Greedy:** `The two are not the same.` **Reranking:** `A "anti-Semite" is someone who does not like Jews or other people of the same religion.`

Figure 4.2: Examples of effective re-ranking to pick contextually specific sequences matching the gold level.

Elaboration Generation Evaluation. Corpus BLEU scores themselves are fairly low, as seen in Table 4.2 and 4.3. However, during manual evaluation of these sequences, we find that elaborations produced after finetuning GPT-2 can be semantically plausible, coherent, and elaboration-like. Content that is pertinent and new, but that does not overlap with the content in the gold elaboration is not rewarded. In some cases, staying true to the content of the gold elaboration is likely unnecessary, as long as the contextual specificity is comparable. To that end, human evaluation of elaboration generation should be emphasized, given that the purpose of elaborations is largely to make simplified documents easier to understand.

Retrieval. Elaborations of medium contextual specificity often involve knowledge not readily available from the simplified or original text. For example, generating factually correct details about a certain event or entity can prove a challenging feat for pre-trained language models. To that end, to generate truly effective elaborations of medium contextual specificity, some type of retrieval module may be necessary.



a No Finetuning

b Simple Corpus Finetuning

Figure 4.3: Predicted contextual specificity distributions on test set for both no-finetuning and simple-finetuning greedy decoding settings.

Distribution of contextual specificity without finetuning. To understand whether GPT-2 is able to generate a range of contextually specific elaborations without being finetuned, we elicit elaborations from GPT-2 and pass these generated elaborations to our best contextual specificity model. Figure 4.3a shows this distribution, compared with the gold labels from our test set for each level of contextual specificity. In addition, we display the same comparison after finetuning GPT-2 on our simple document corpus, as shown in 4.3b. We see that before finetuning, GPT-2 is able to generate a distribution of contextually specific elaborations fairly similar to the gold. However, the disparity between classes grows after finetuning – most notably, after finetuning, GPT-2 seems to generate majority high contextualization elaborations, according to our best model from Section 4.1.

4.3 Summary

In this chapter, we presented two tasks – contextual specificity prediction and elaboration generation, and established baselines for both tasks using large scale pre-trained LMs. We explored the role of document context from both the original and simplified documents. Our highest performing contextual specificity model achieved an F1 score of 53.6, indicating plenty of room for improvement to build *reliable* contextualization models that could prove useful downstream during elaboration generation. Likewise, while our generation models did generate some coherent, elaboration-like sentences, they did not perform exceedingly well, indicating the need for better techniques, likely involving an approach that uses content from the original document in a natural way, serving as a source for possible context.

Chapter 5

Conclusion and Future Work

5.1 Summary

As we move from sentence simplification towards automatic document simplification, it is important to understand what makes simplifying a document fundamentally different from simplifying a standalone sentence. In this work, we presented a new phenomenon we observe in document simplification we call elaborative simplification, involving the insertion of content to make simplified texts easier to understand. Our primary motivation was to understand *how* entities, ideas, or concepts are elaborated during the process of simplification. We discussed construction of a new corpus of over 1.3K verified elaborations through a combination of automatic processing and human verification. Using this corpus, we qualitatively analyzed elaborations, introducing a new framework to study the manner in which things are elaborated – contextual specificity. Using this framework, we proposed two new modeling tasks related to elaborative simplification – contextual specificity prediction, and elaboration generation, and established baselines for both using large scale pre-trained language models.

5.2 Future Work

There are ample directions for future work given this initial study of elaborative simplification. First, the Newsela simplification corpus contains several articles per document set, ranging across multiple grade levels. This could present an interesting opportunity to understand how elaborative simplification might change when generated for different audiences at varying reading levels, similar to the controllable sentence simplification system developed by Martin et al. (2019).

Second, to streamline elaboration detection in simplified texts, better alignment algorithms could potentially decrease the amount of human verification needed to collect elaborations at scale. Collecting a larger corpus of elaborations would not only allow for more stable finetuning of pre-trained language models, but would also allow for larger scale analyses of elaborative simplification.

One of the main sources of *relevant* content for a given document set is the original document. When crafting our model inputs, we simply concatenated text from the original document and simplified document, and separated them via a special token. However, this method led to a drop in performance. One direction for future work is better incorporation of text or information from the original document for both contextual specificity prediction and elaboration generation. One possible approach is to elicit information from LMs regarding content in the original document as an intermediate step, similar to Shwartz et al. (2020). Another is to pre-process the original document in some capacity to extract relevant information.

To utilize elaborative simplification systems in practical settings, for example during real-time simplification, a reliable model to predict level of contextual specificity from input context alone is necessary. Our best performing contextual specificity prediction model based on context alone achieved an F1 score of 44.9 on the test set, indicating ample room for improvement.

Our current re-ranking system for candidate generated elaborations consists of relying on contextual specificity alone. While this did show an improvement, we found that generated sequences were relevant only to a small snippet of context before the gold elaboration. When evaluated as a part of the whole document, many elaborations contained irrelevant information. One direction for future work is to augment reranking parameters – i.e evaluate generated sequences for relevance to the original document content, in addition to contextual specificity.

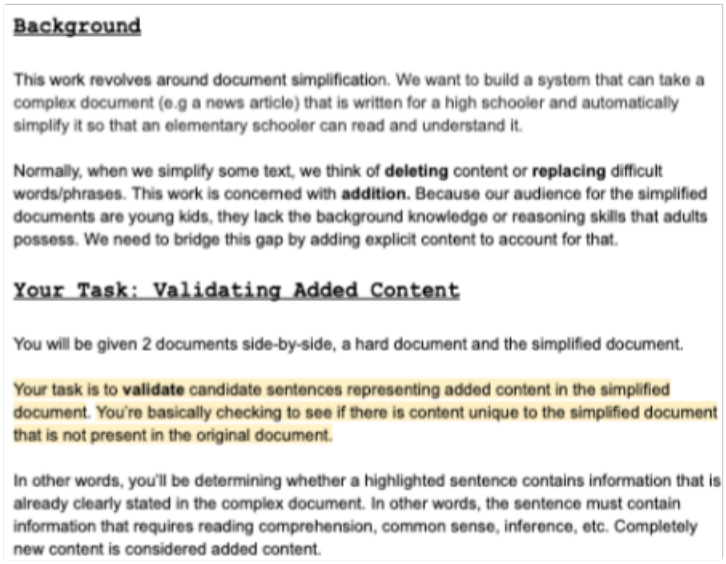
Depending on the kind of elaborations one may want to generate, one possible direction for future investigation is to finetune elaboration generation systems on different datasets ranging in contextual specificity. For example, if low contextualization elaborations are routinely desired, then it is possible to train on an entity post-modifier dataset, such as in Kang et al. (2019). This could potentially increase training resource size and lead to improved stability in results.

We presented one phenomenon in document simplification we empirically observed in the Newsela simplification corpus. Elaborative simplification is one among many phenomena in document simplification. The identification and study of other phenomena is essential for developing effective and accurate document simplification systems.

Lastly, our work involved studying the Newsela corpus, a corpus of news articles. We observed elaborative simplification as a phenomenon during the simplification of news articles, however one future direction of work is to understand elaborative simplification in different domains.

Appendix A

Supplemental Material



Background

This work revolves around document simplification. We want to build a system that can take a complex document (e.g a news article) that is written for a high schooler and automatically simplify it so that an elementary schooler can read and understand it.

Normally, when we simplify some text, we think of **deleting** content or **replacing** difficult words/phrases. This work is concerned with **addition.** Because our audience for the simplified documents are young kids, they lack the background knowledge or reasoning skills that adults possess. We need to bridge this gap by adding explicit content to account for that.

Your Task: Validating Added Content

You will be given 2 documents side-by-side, a hard document and the simplified document.

Your task is to **validate** candidate sentences representing added content in the simplified document. You're basically checking to see if there is content unique to the simplified document that is not present in the original document.

In other words, you'll be determining whether a highlighted sentence contains information that is already clearly stated in the complex document. In other words, the sentence must contain information that requires reading comprehension, common sense, inference, etc. Completely new content is considered added content.

Figure A.1: Sample overview provided to expert annotators.

.

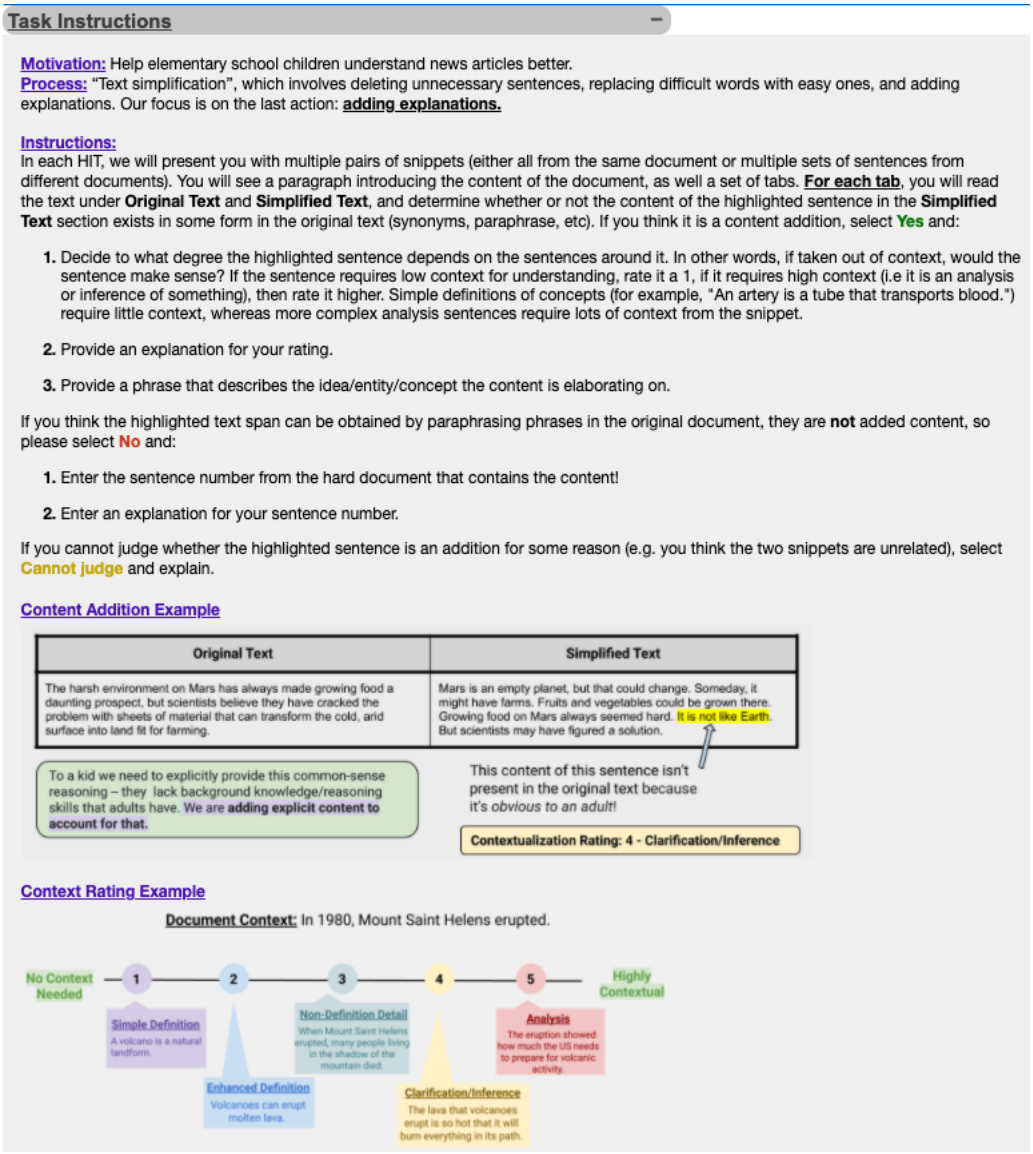


Figure A.2: Annotation instructions provided to crowdworkers.

.

Bibliography

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, 2017.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1): 135–187, 2020.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. 2016.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10, 1998.

John A Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. Simplifying text for language-impaired readers. In *Ninth Confer-*

ence of the European Chapter of the Association for Computational Linguistics, 1999.

Han-Bin Chen, Hen-Hsen Huang, Hsin-Hsi Chen, and Ching-Ting Tan. A simplification-translation-restoration framework for cross-domain smt applications. In *Proceedings of COLING*, pages 545–560, 2012.

William Coster and David Kauchak. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, 2011.

Scott A Crossley, Max M Louwerse, Philip M McCarthy, and Danielle S McNamara. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30, 2007.

Joe Davison, Joshua Feldman, and Alexander M Rush. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, 2019.

Jan De Belder and Marie-Francine Moens. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26, 2010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. Editnts:

An neural programmer-interpreter model for sentence simplification through explicit editing. *arXiv preprint arXiv:1906.08104*, 2019.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. Large-scale qa-srl parsing. *arXiv preprint arXiv:1805.05377*, 2018.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, 2015.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73, 2013.

Jun Seok Kang, Robert L Logan IV, Zewei Chu, Yang Chen, Dheeru Dua, Kevin Gimpel, Sameer Singh, and Niranjan Balasubramanian. Pomo: Generating entity-specific post-modifiers in context. *arXiv preprint arXiv:1904.03111*, 2019.

Klaus Krippendorff. Reliability. *The International Encyclopedia of Communication*, 2008. doi: 10.1002/9781405186407.wbiecr029.

Reno Kriz, Joao Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. Complexity-weighted loss and diverse reranking for sentence simplification. *arXiv preprint arXiv:1904.02767*, 2019.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. Iterative edit-based unsupervised sentence simplification. *arXiv preprint arXiv:2006.09639*, 2020.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema

challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Angrosh Annayappan Mandya, Tadashi Nomoto, and Advaith Siddharthan. Lexico-syntactic text simplification and compression with typed dependencies. In *25th International Conference on Computational Linguistics*, 2014.

Louis Martin, Benoît Sagot, Eric de la Clergerie, and Antoine Bordes. Controllable sentence simplification. *arXiv preprint arXiv:1910.02677*, 2019.

Jana M Mason. Facilitating reading comprehension through text structure manipulation. *Center for the Study of Reading Technical Report; no. 092*, 1978.

Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. Why is that relevant? collecting annotator rationales for relevance judgments. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

Shashi Narayan, Claire Gardent, Shay B Cohen, and Anastasia Shimorina. Split and rephrase. *arXiv preprint arXiv:1707.06971*, 2017.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, 2017.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical

literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access, 2018.

Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. Massalign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4, 2017.

Gustavo Henrique Paetzold. *Lexical Simplification for Non-Native English Speakers*. PhD thesis, University of Sheffield, 2016.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*, 2017.

Sheena Panthaplackel, Pengyu Nie, Milos Gligoric, Junyi Jessy Li, and Raymond J Mooney. Learning to update natural language comments based on code changes. *arXiv preprint arXiv:2004.12169*, 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Rebecca Passonneau. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. 2006.

David Pellow and Maxine Eskenazi. An open corpus of everyday documents for simplification tasks. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 84–93, 2014.

Sarah E Petersen and Mari Ostendorf. Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology in Education*, 2007.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219, 2013.

Steven Ross, Michael H Long, and Yasukata Yano. Simplification or elaboration? the effects of two types of text modifications on foreign language reading comprehension. *University of Hawai'i Working Papers in English as a Second Language 10 (2)*, 1991.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. *arXiv preprint arXiv:2004.05483*, 2020.

Advaith Siddharthan. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298, 2014.

Sara Botelho Silveira and António Branco. Enhancing multi-document summaries with sentence simplification. In *Proceedings on the International Conference on Artificial Intelligence*, page 1, 2012.

Sanja Štajner and Maja Popović. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242, 2016.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618, 2007.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910, 2019.

Kristian Woodsend and Mirella Lapata. Wikisimple: Automatic simplification of wikipedia articles. In *Proceedings of the 25th National Conference on Artificial Intelligence*, pages 927–932, San Francisco, CA, 2011.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.

Yasukata Yano, Michael H Long, and Steven Ross. The effects of simplified and elaborated texts on foreign language reading comprehension. *Language learning*, 44(2):189–219, 1994.

Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*, 2017.

Sanqiang Zhao, Rui Meng, Daqing He, Saptono Andi, and Parmanto Bambang. Integrating transformer and paraphrase rules for sentence simplification. *arXiv preprint arXiv:1810.11193*, 2018.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. Discourse level factors for sentence deletion in text simplification. *arXiv preprint arXiv:1911.10384*, 2019.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, 2010.